

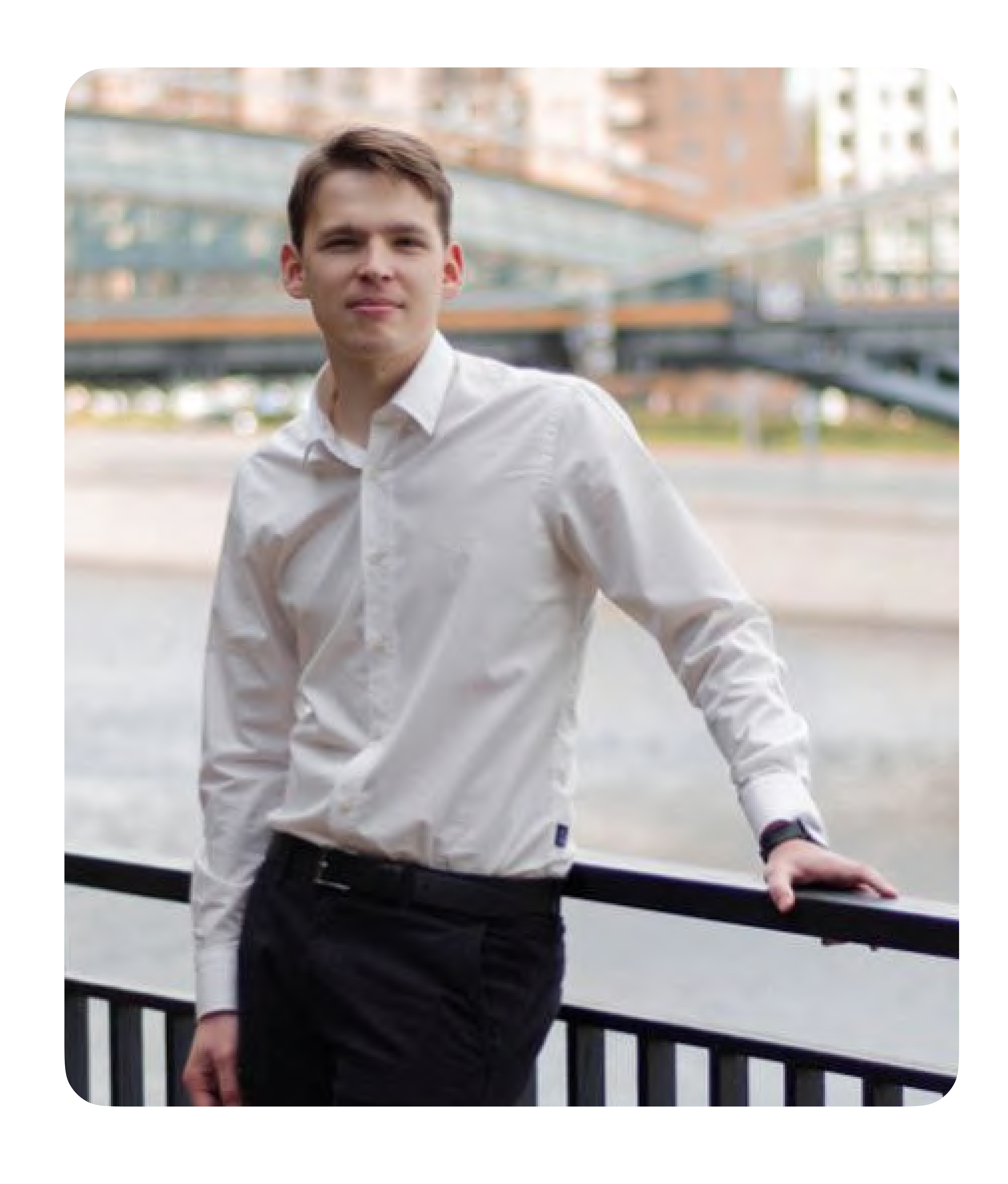
Высокопроизводительный промышленный сервис для компьютерного зрения на Python

Алексеенко Григорий, Layer SberDevices



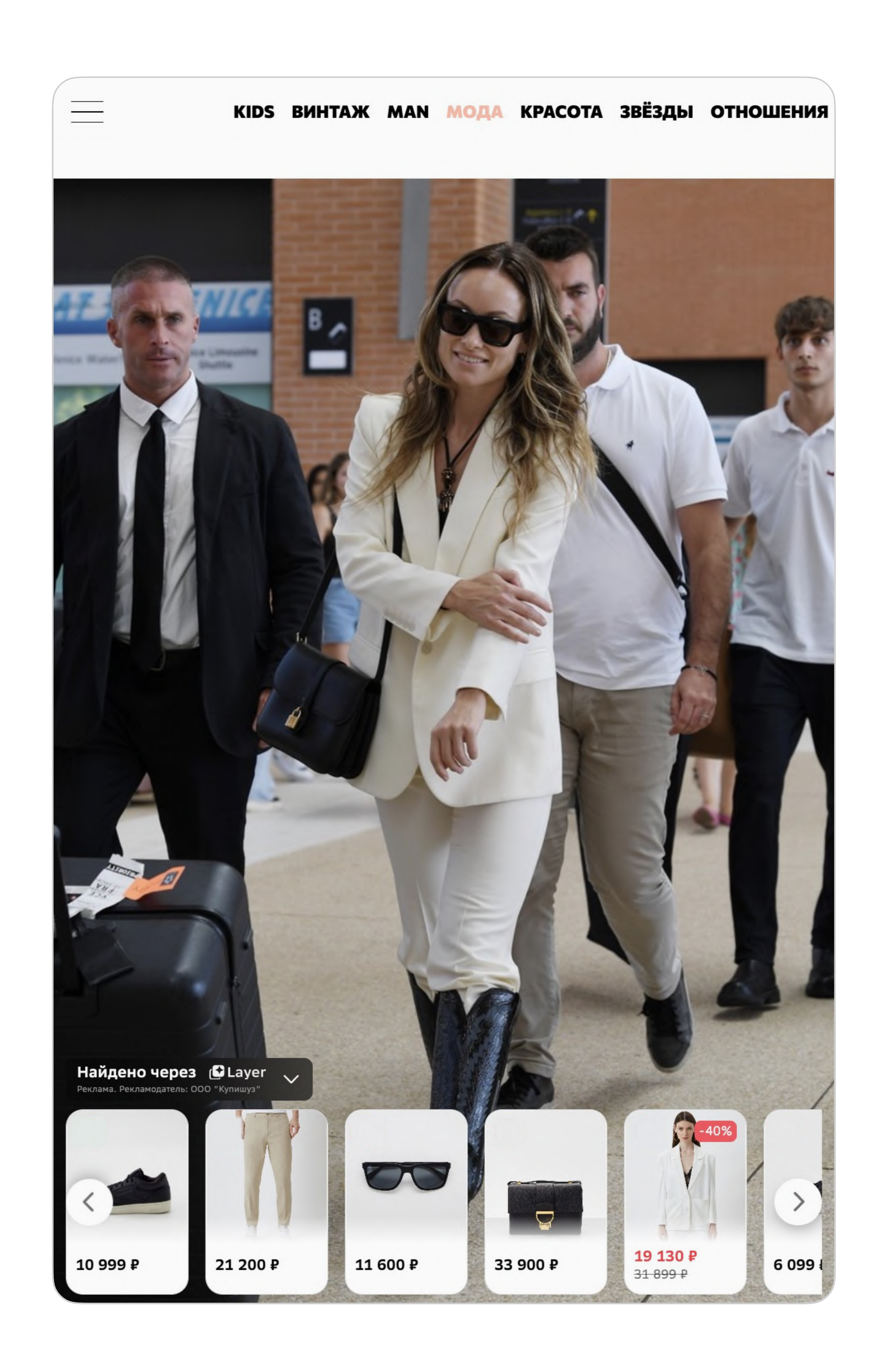


Немного о себе



1. Senior Data Scientist в Layer с 2021
2. Работаю в компьютерном зрении пятый год
3. Пишу на Python и C++
4. Занимался различными проектами от видеоаналитики для ритейла и умных городов до подсчета свиней на ферме

Монетизация фото- и видеоконтента: In-Image



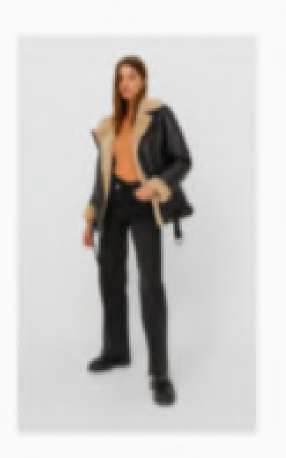


Особенности

1. Каждый день новые статьи
2. Большое количество просмотров, нужно чтобы при просмотре данные были актуальные
3. Декодируются jpeg, png и прочие форматы изображений

Монетизация фото- и видеоконтента: Visual Search

Куртка кожаная женская XS

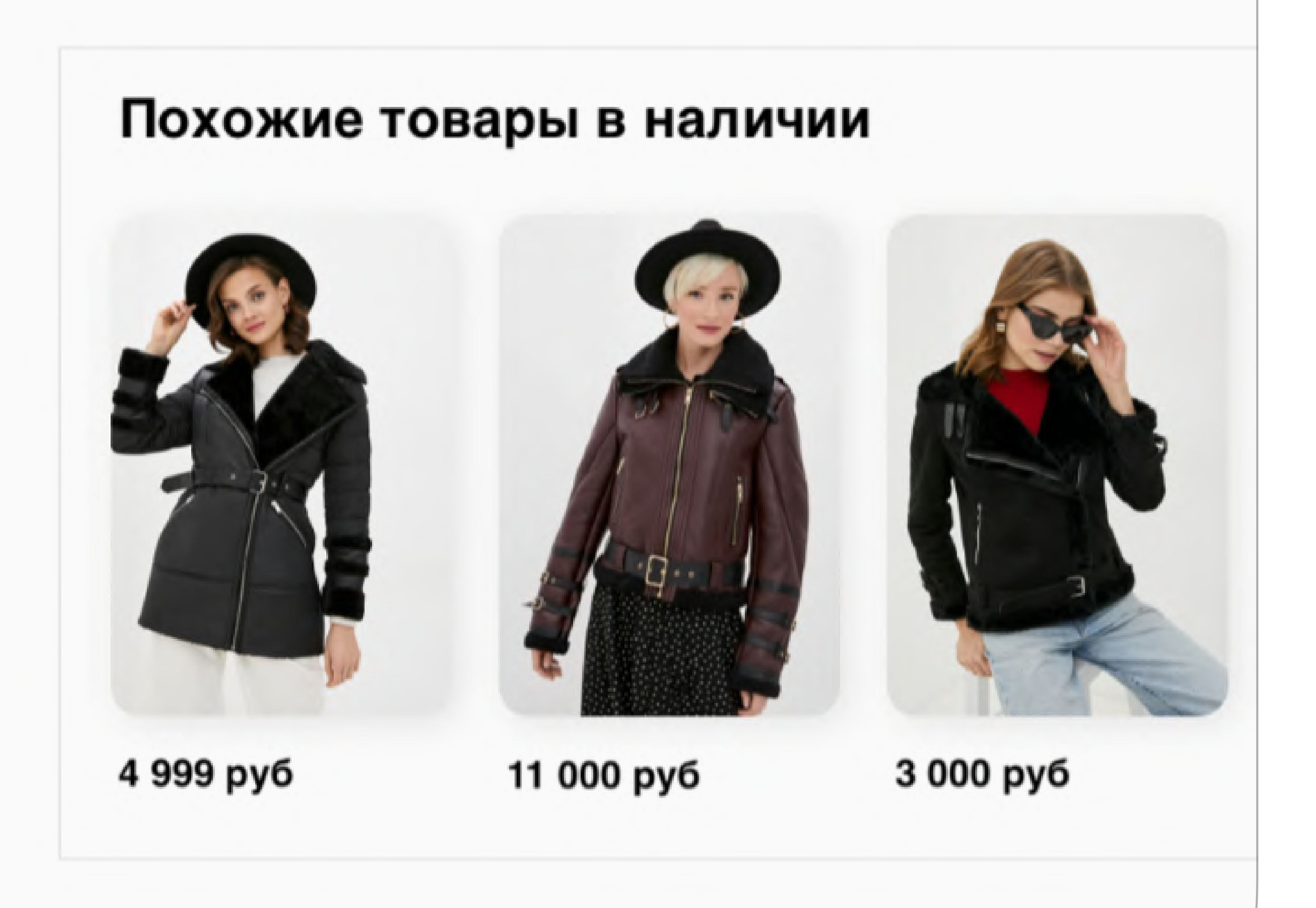






Состав: искусственная кожа
Код товара: 100012302909

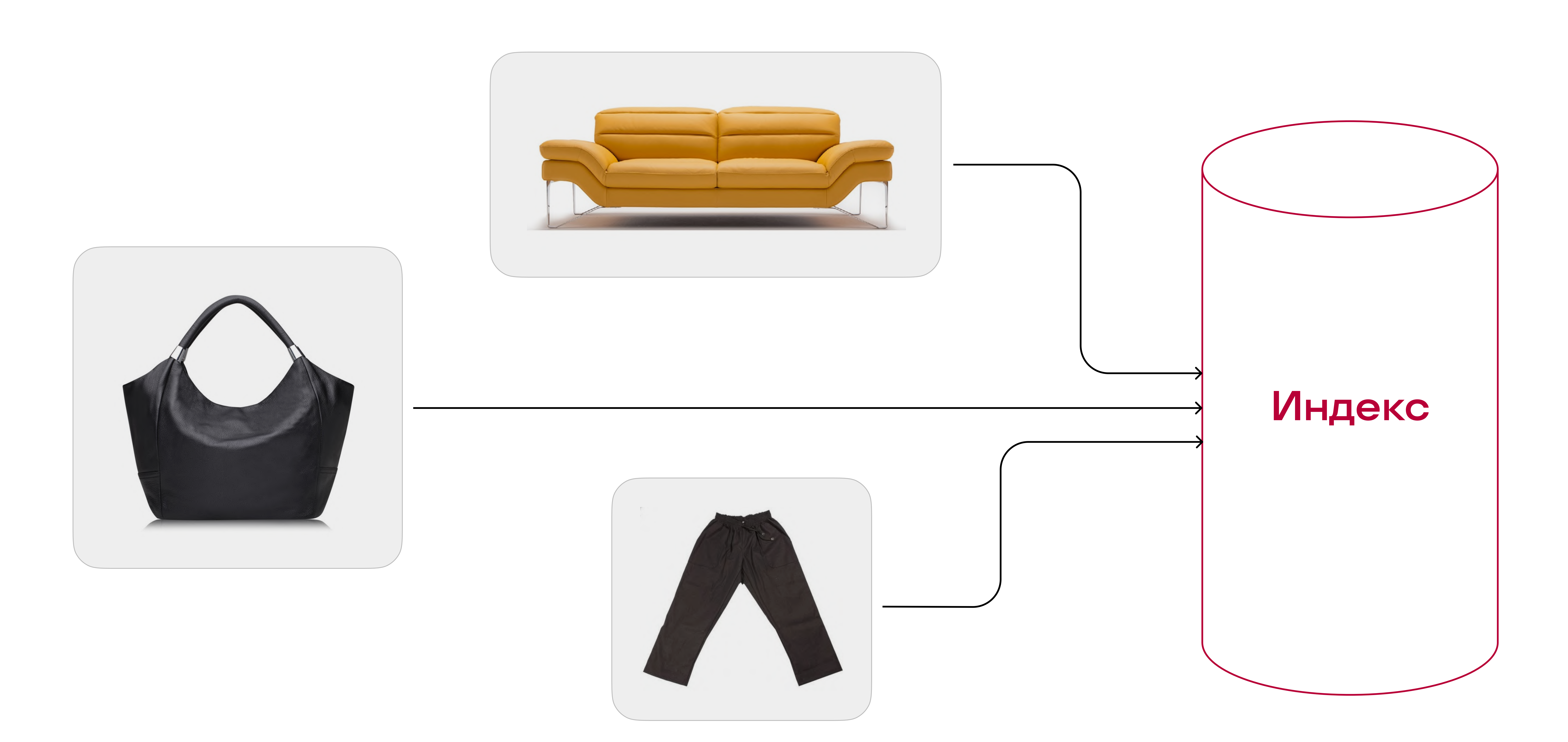
Товар закончился



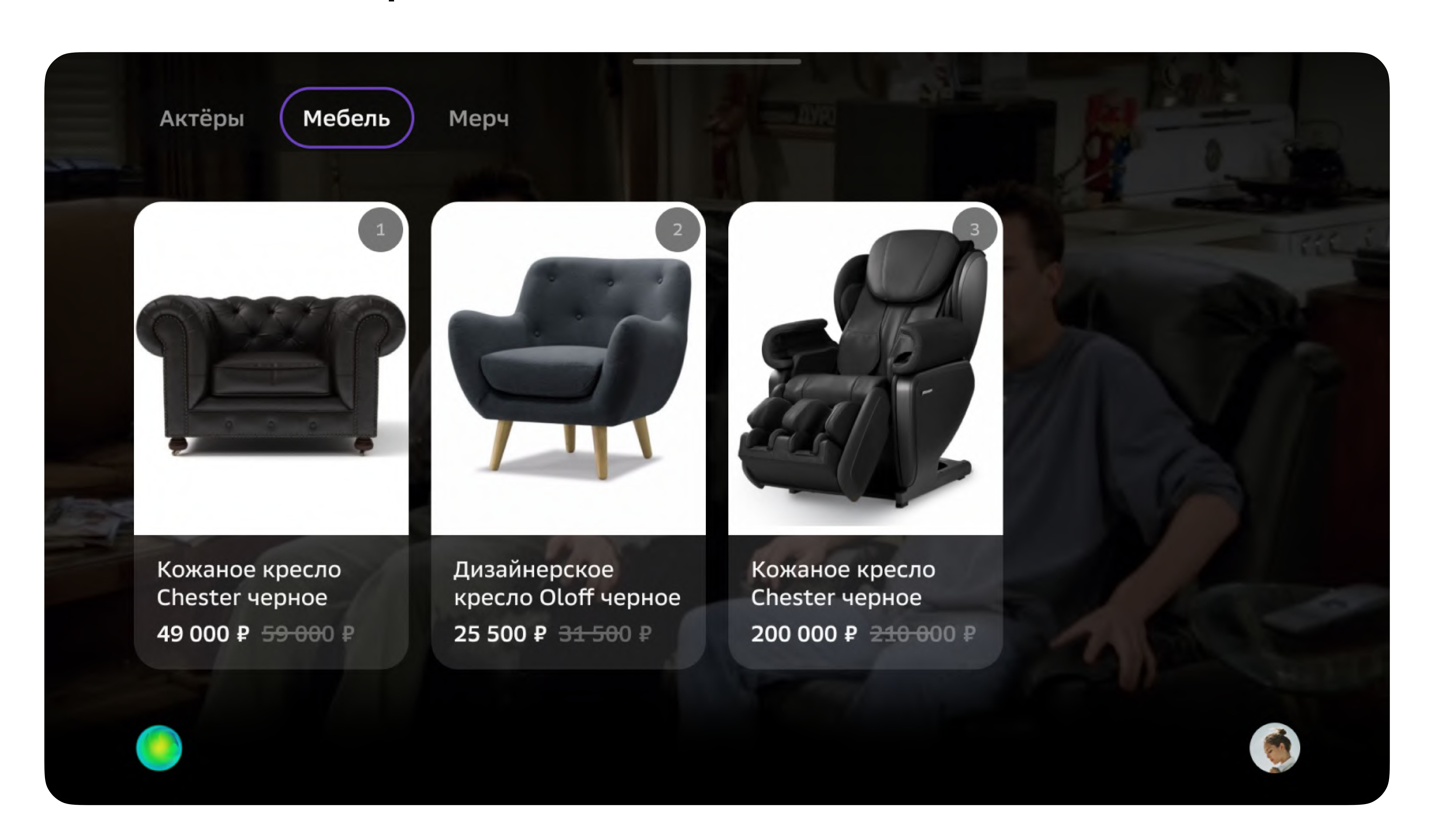
Особенности

1. Миллионы товаров
2. Нужно постоянно обновлять
3. Непредсказуемые форматы изображений
4. Поиск по большому индексу

Монетизация фото- и видеоконтента: генерация индекса



Монетизация фото- и видеоконтента: In-Video

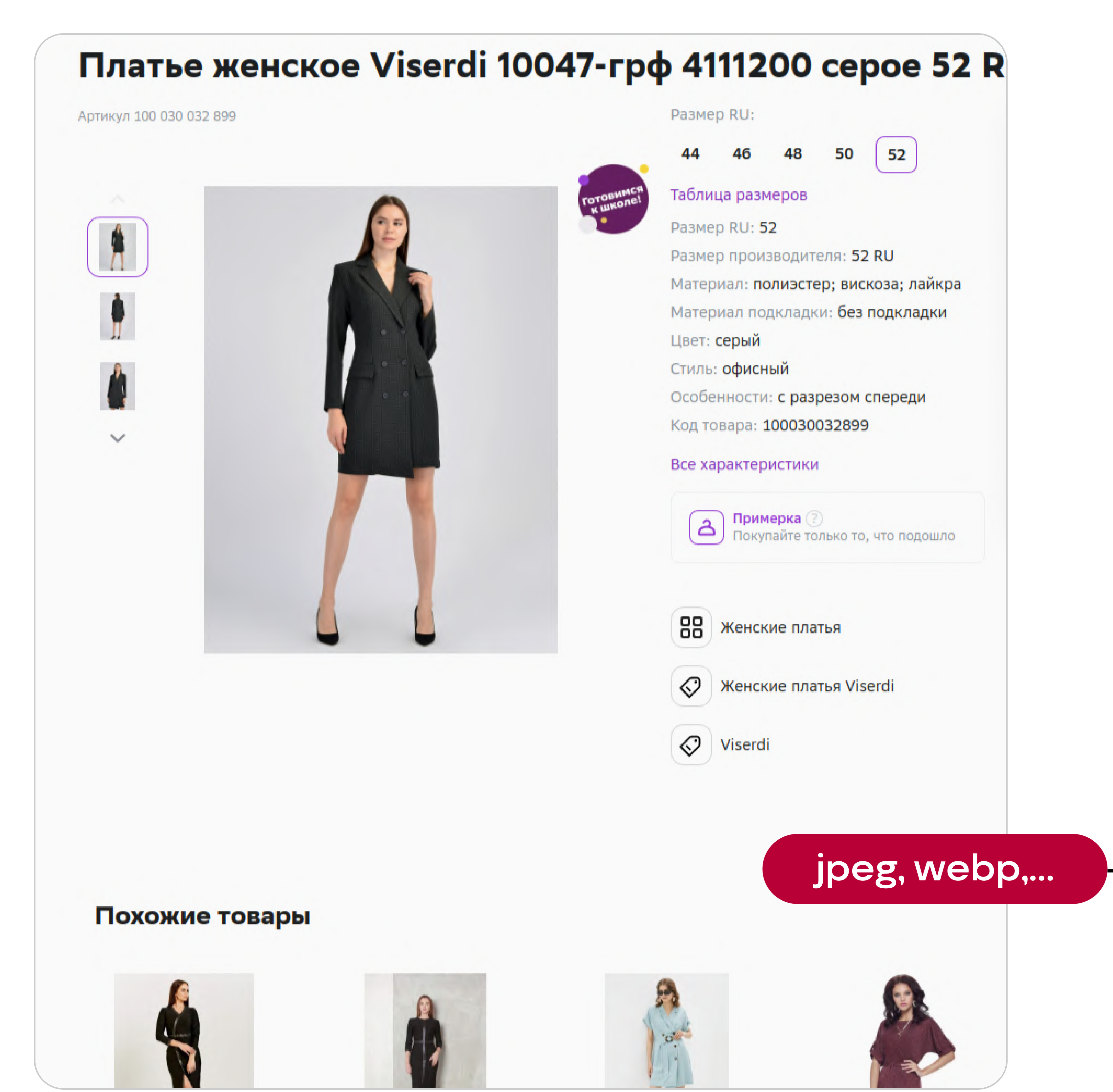


Особенности

1. Работа в реальном времени
2. Максимум 40 мс на ответ !
3. Нужно декодировать видео

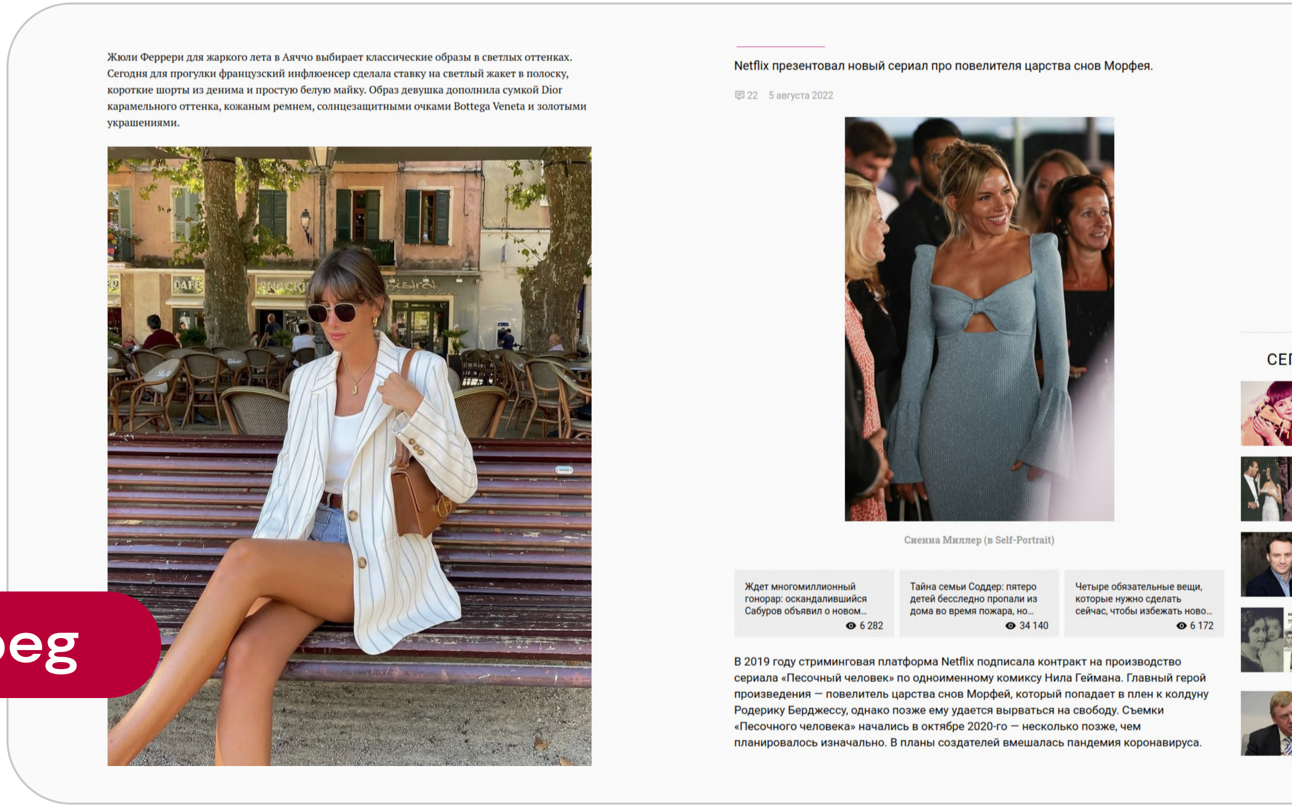
Нагрузки

Visual Search
Генерация индексов



jpeg, webp,...

In-Image



jpeg

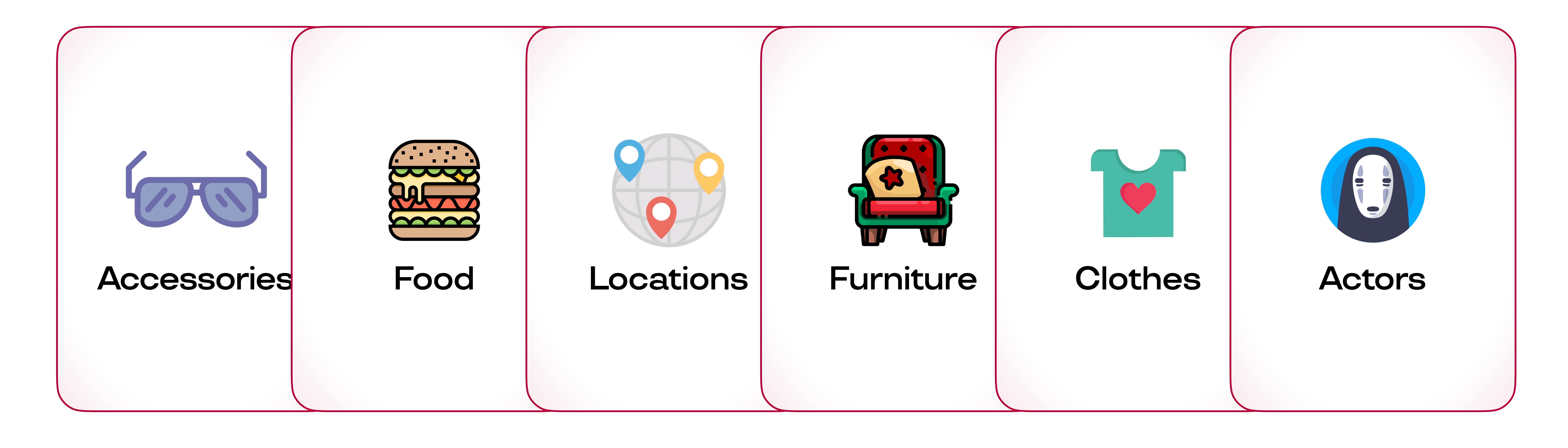
In-Video



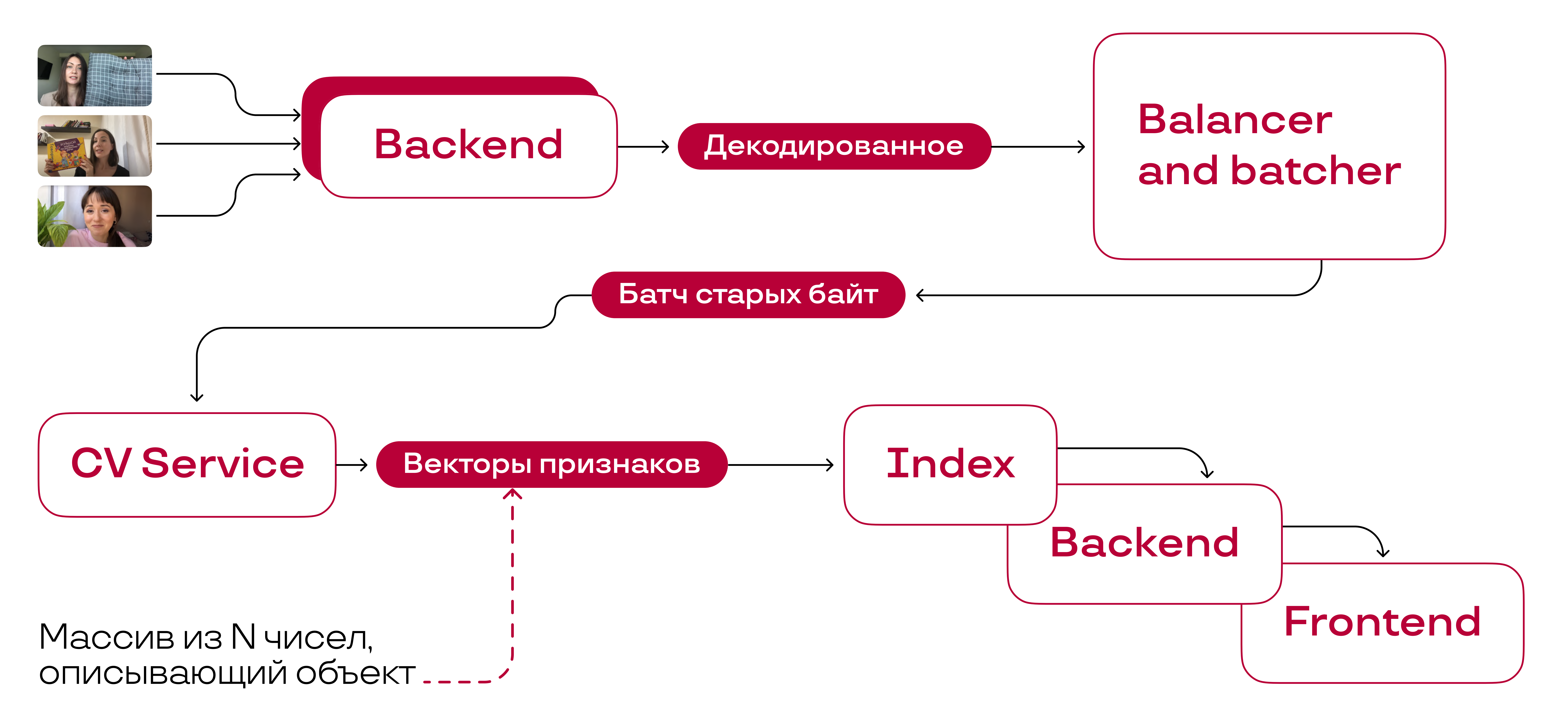
HLS, WebRTC

Категории

Layer



Клиент-серверная архитектура



Детали пайплайна распознавания одежды



sweater
shorts
shoes
shoes

Экстракция объектов

Эмбеддинг

Определение пола и возраста

Пол: **женский**
Возраст: **39**

Поиск по индексу

Варианты оптимизации

Модель

Инференс и батчинг

Вспомогательный код

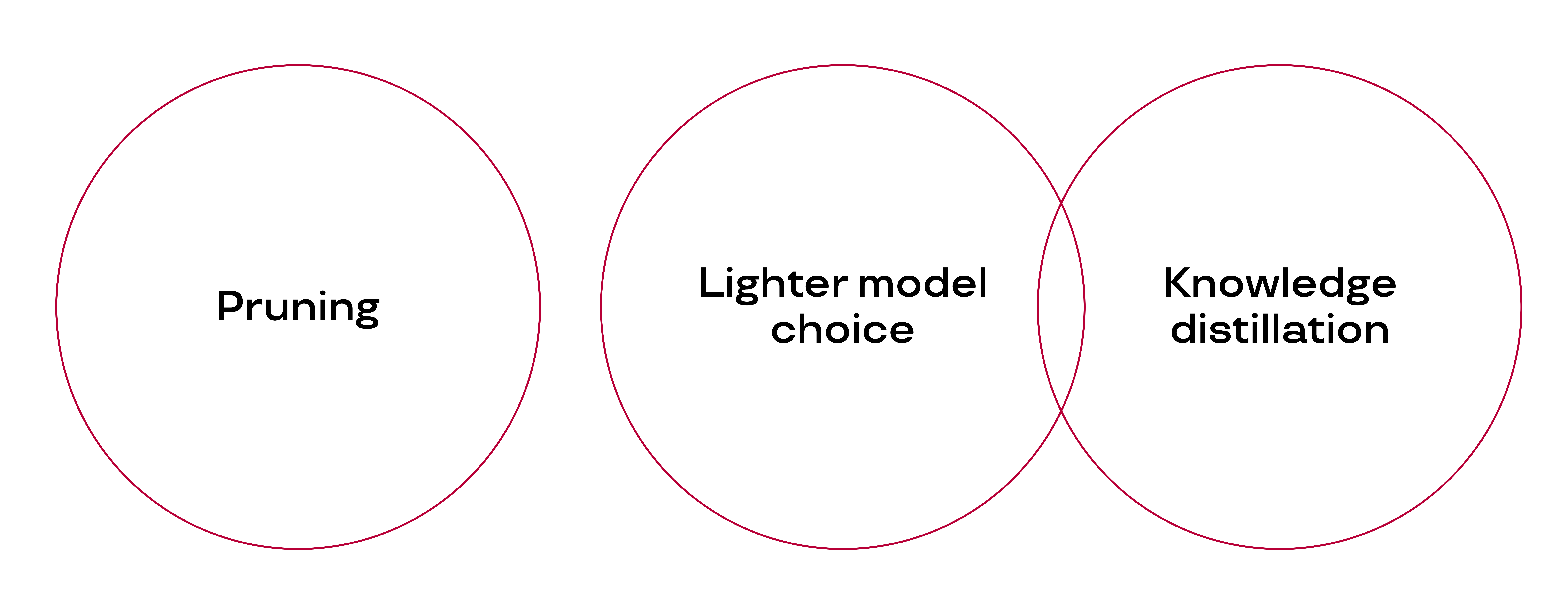
Варианты оптимизации

Модель

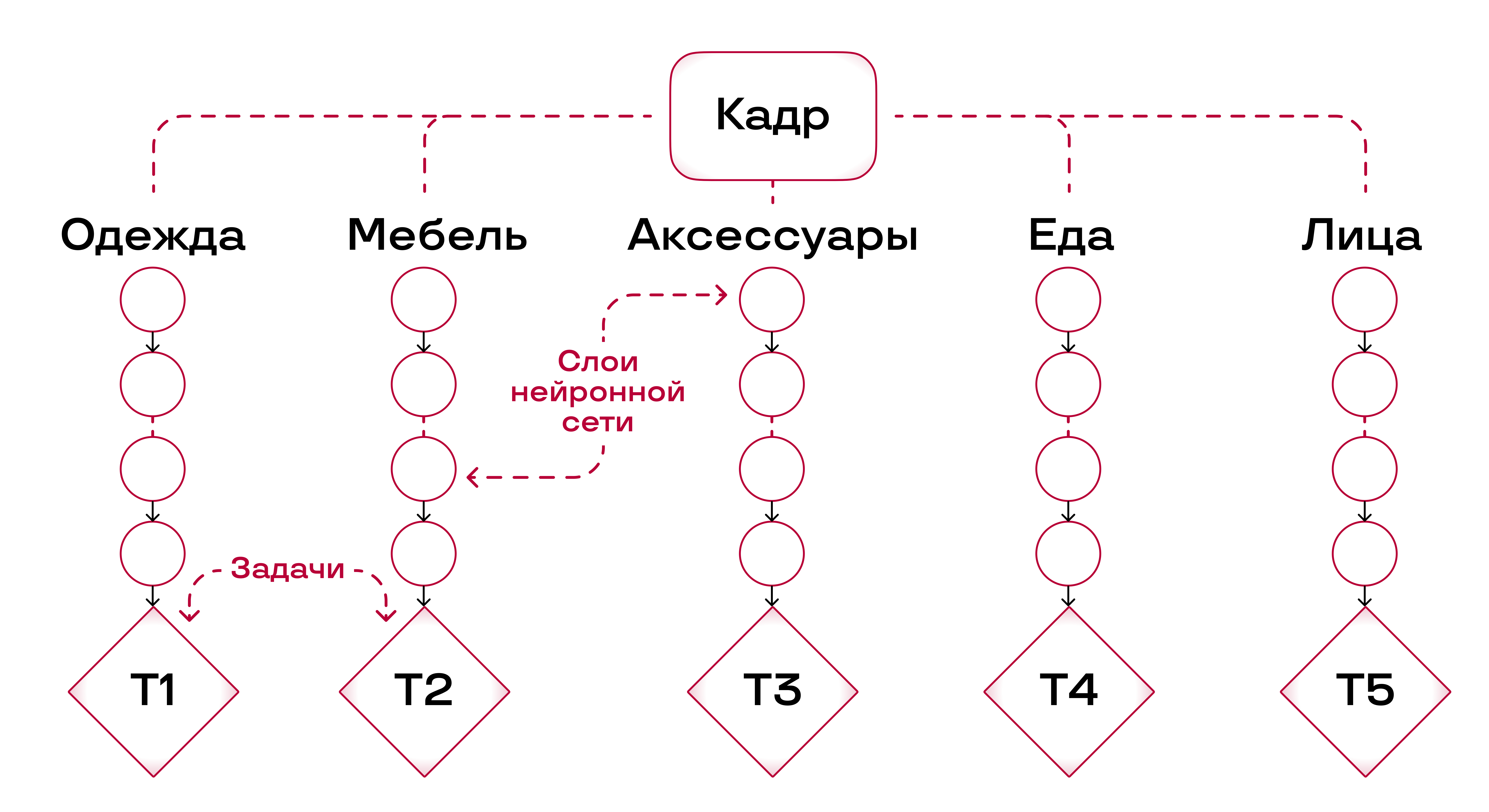
Инференс и батчинг

Вспомогательный код

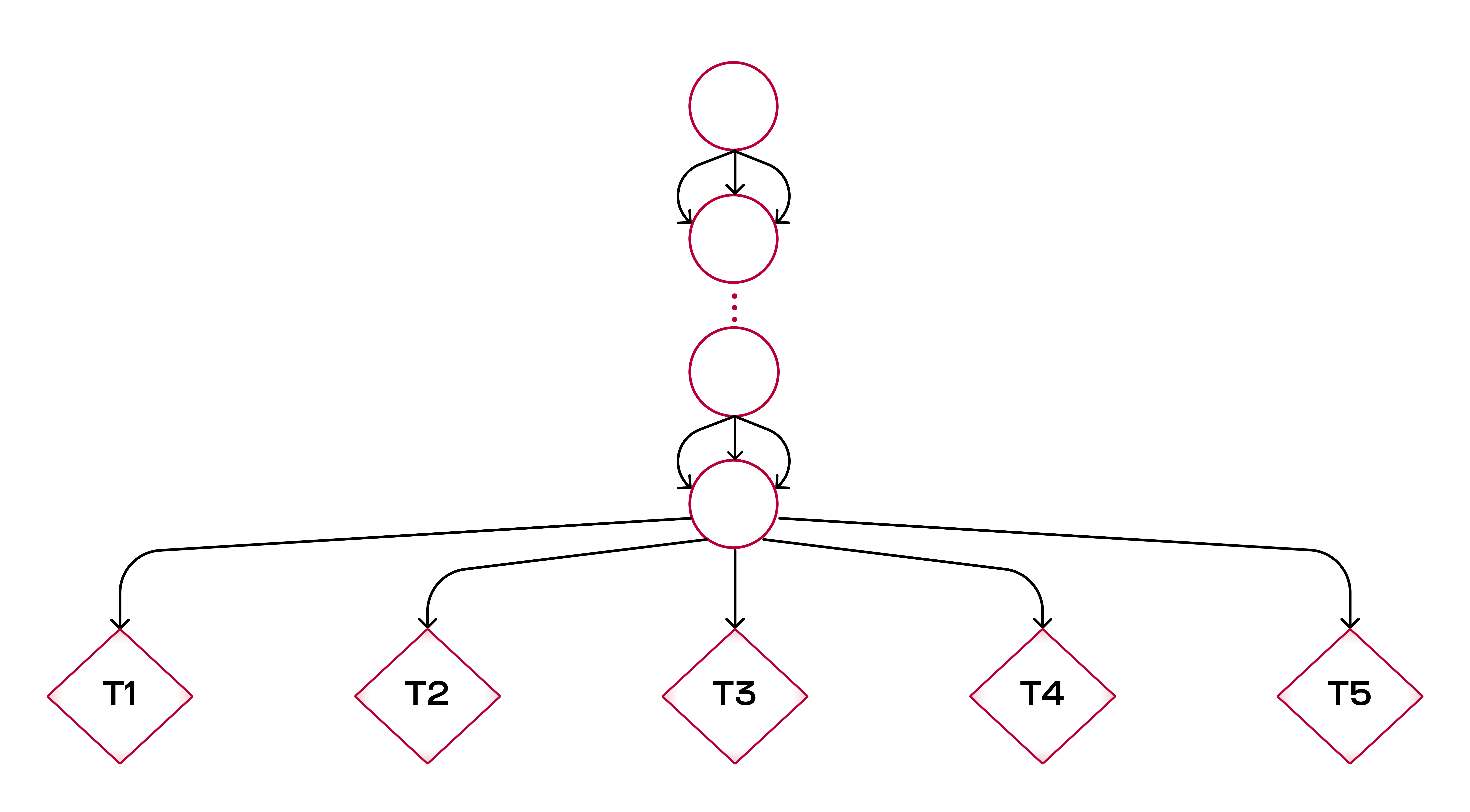
Оптимизация моделей



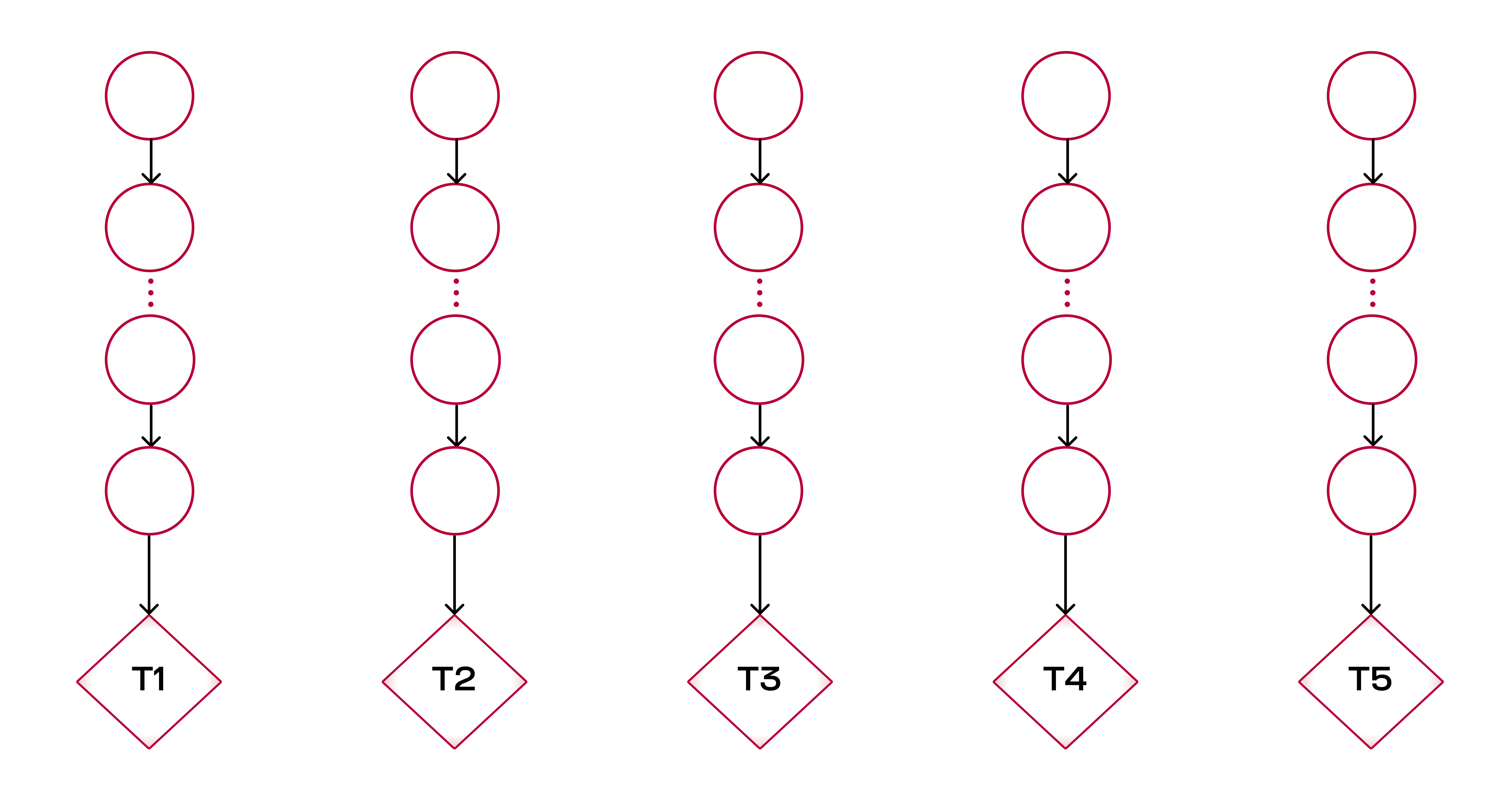
Оптимизация моделей: многозадачность



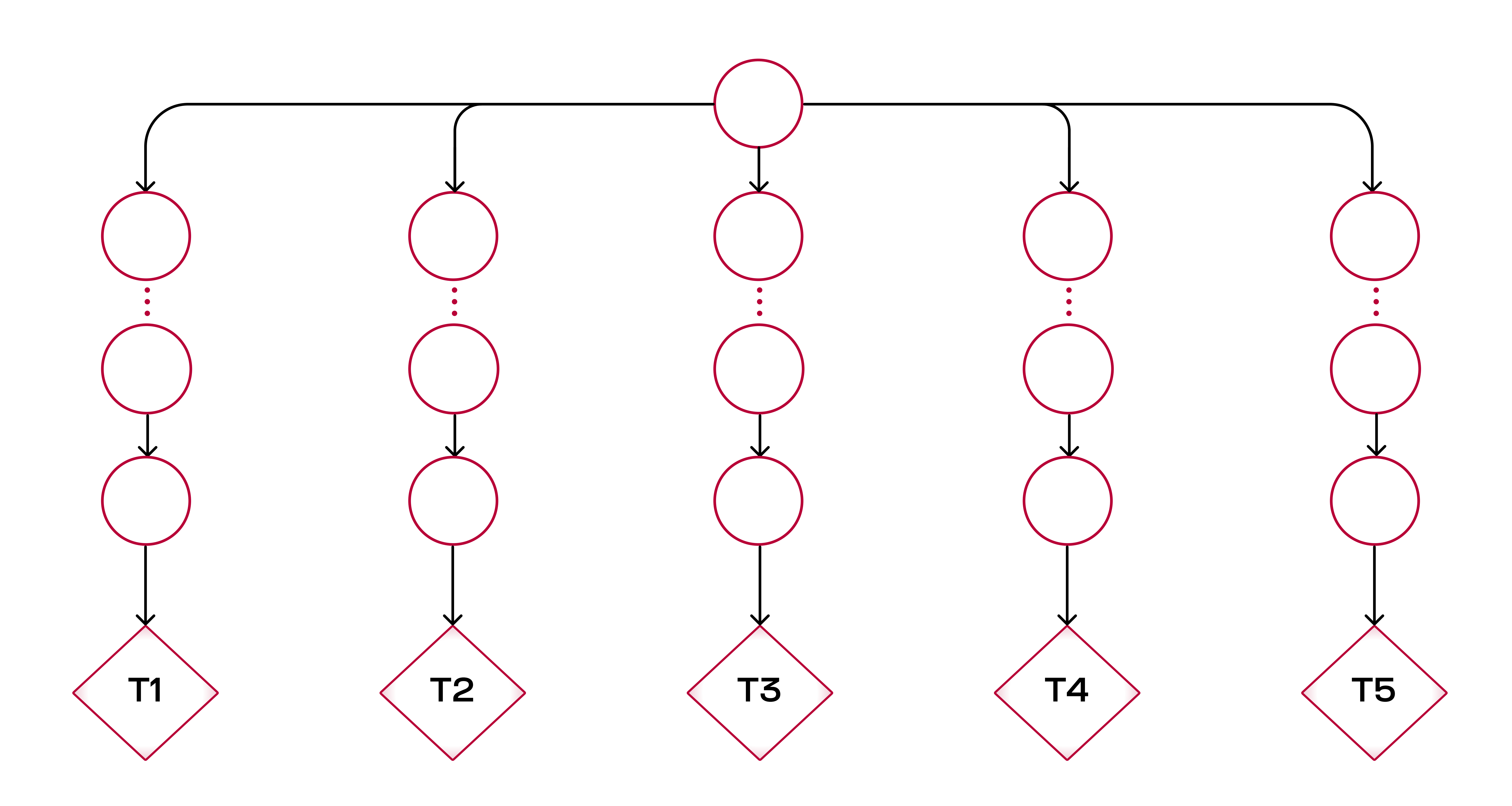
Оптимизация моделей: многозадачность



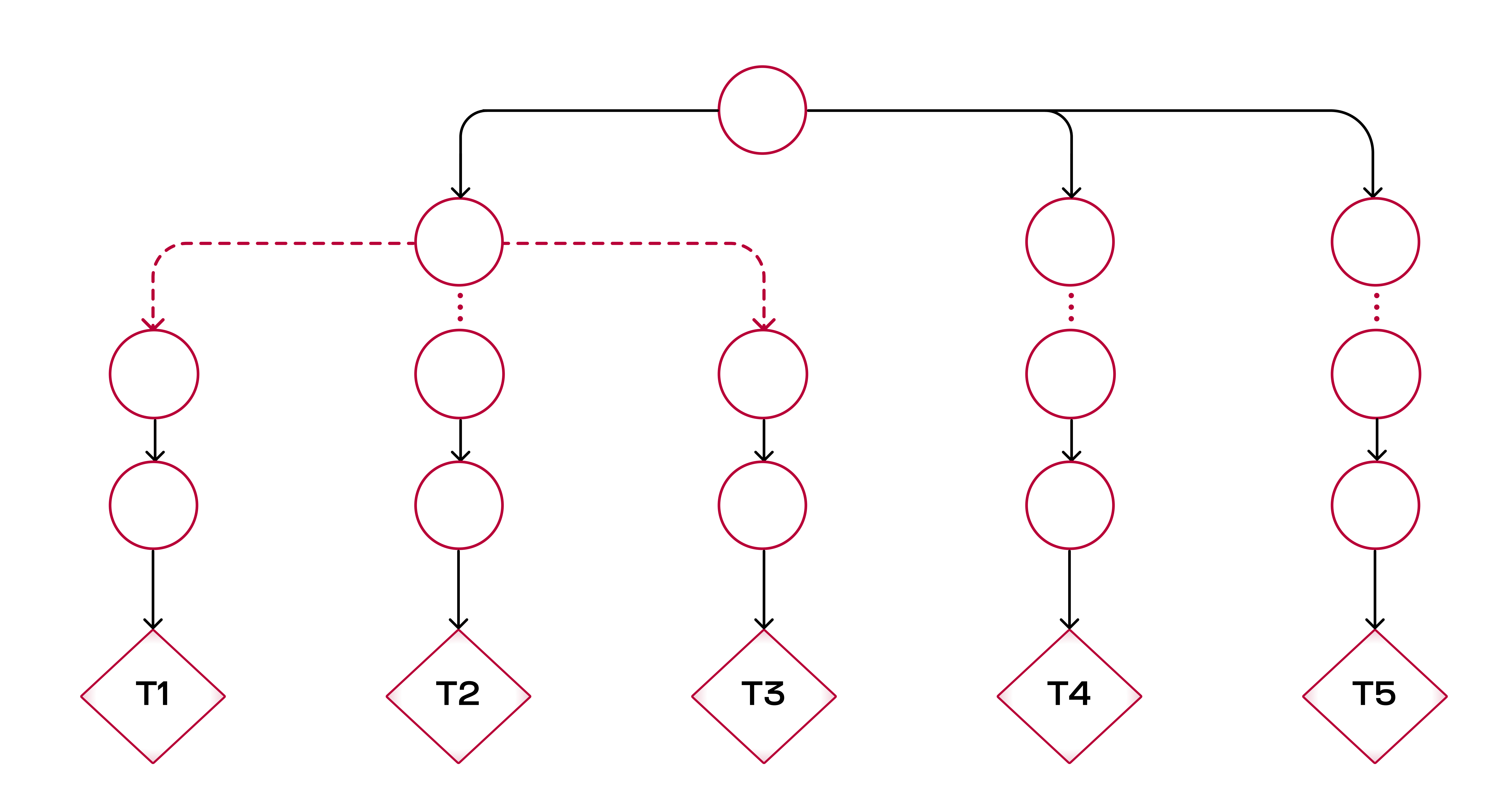
Оптимизация моделей: многозадачность



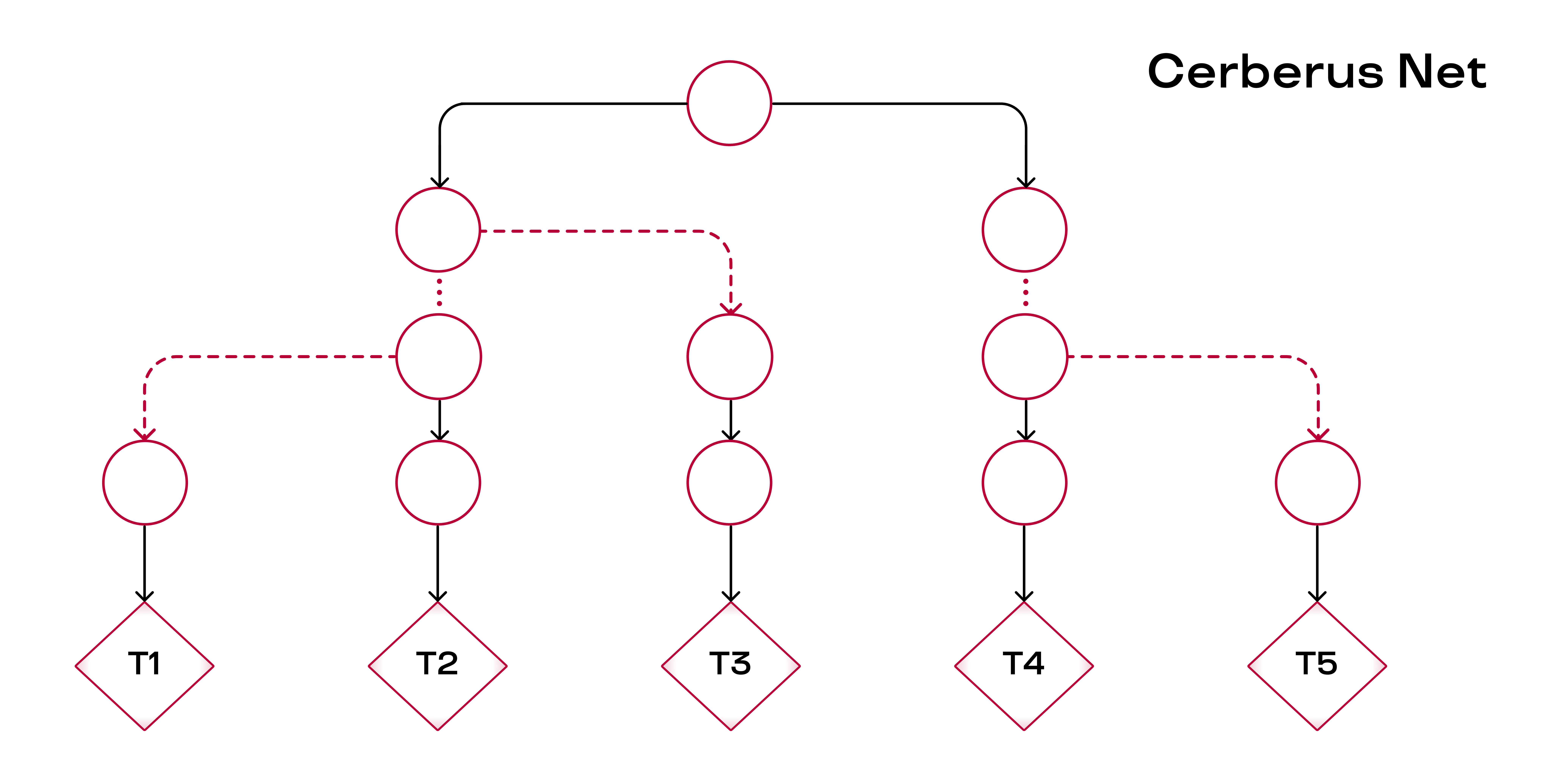
Оптимизация моделей: многозадачность



Оптимизация моделей: многозадачность



Оптимизация моделей: многозадачность



Варианты оптимизации

Модель

Инференс и батчинг

Вспомогательный код

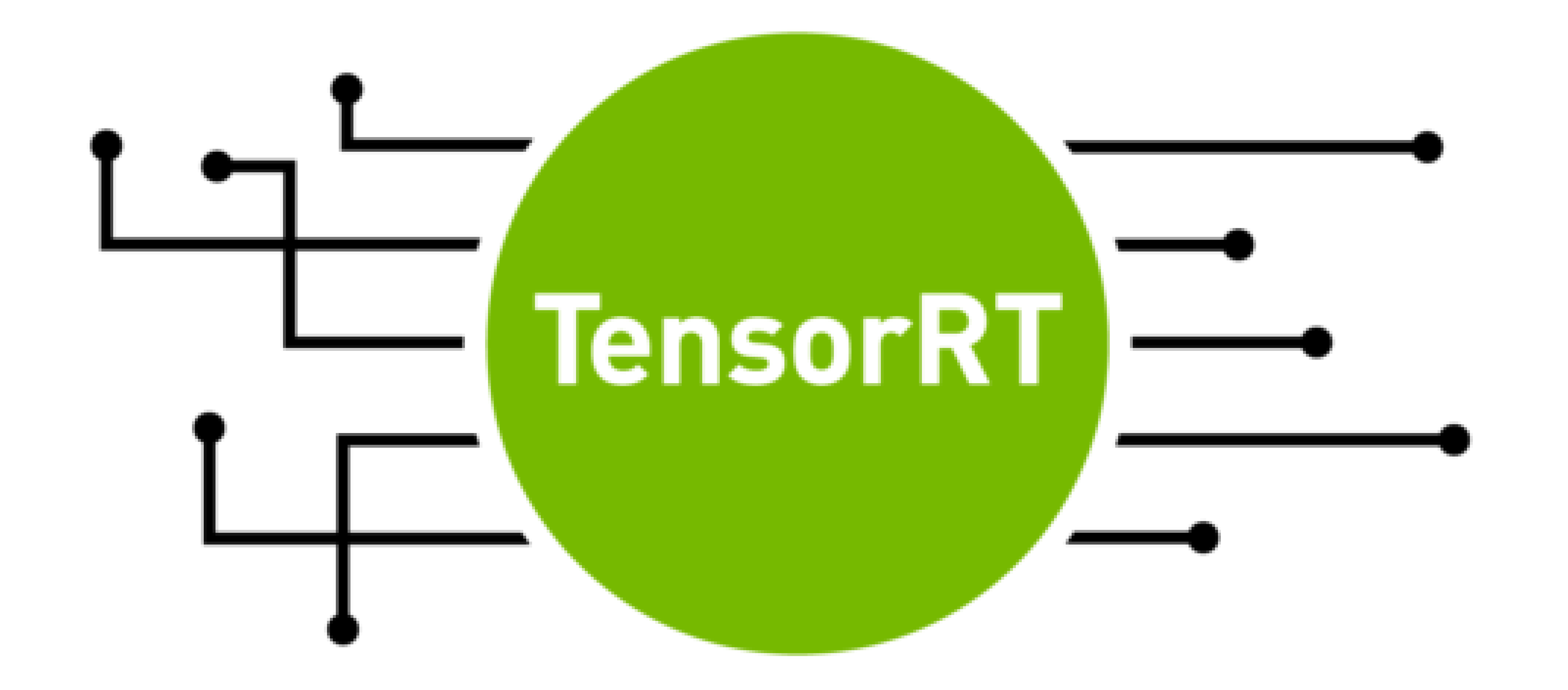
Варианты оптимизации

Модель

Инференс и батчинг

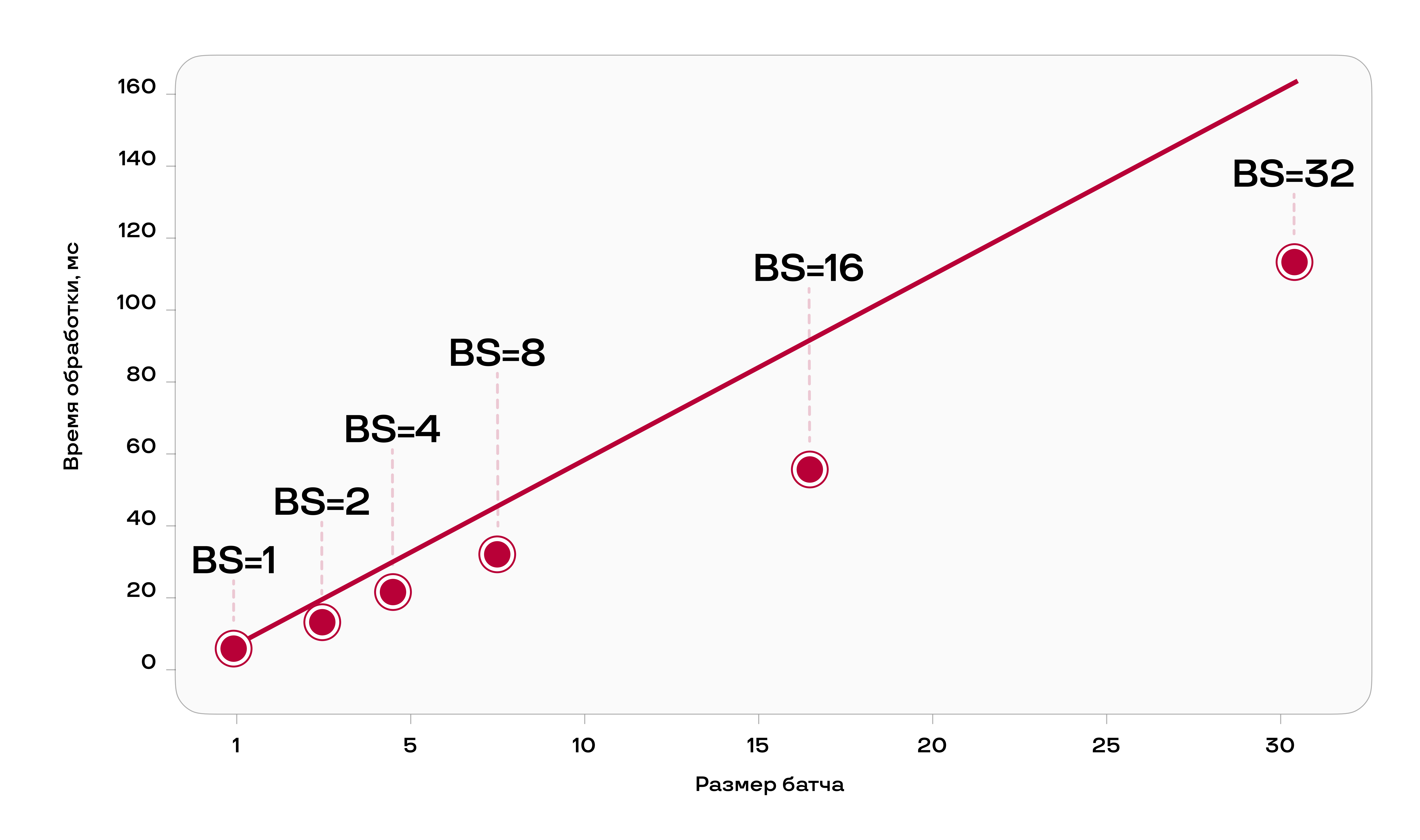
Вспомогательный код

На что обратить внимание:

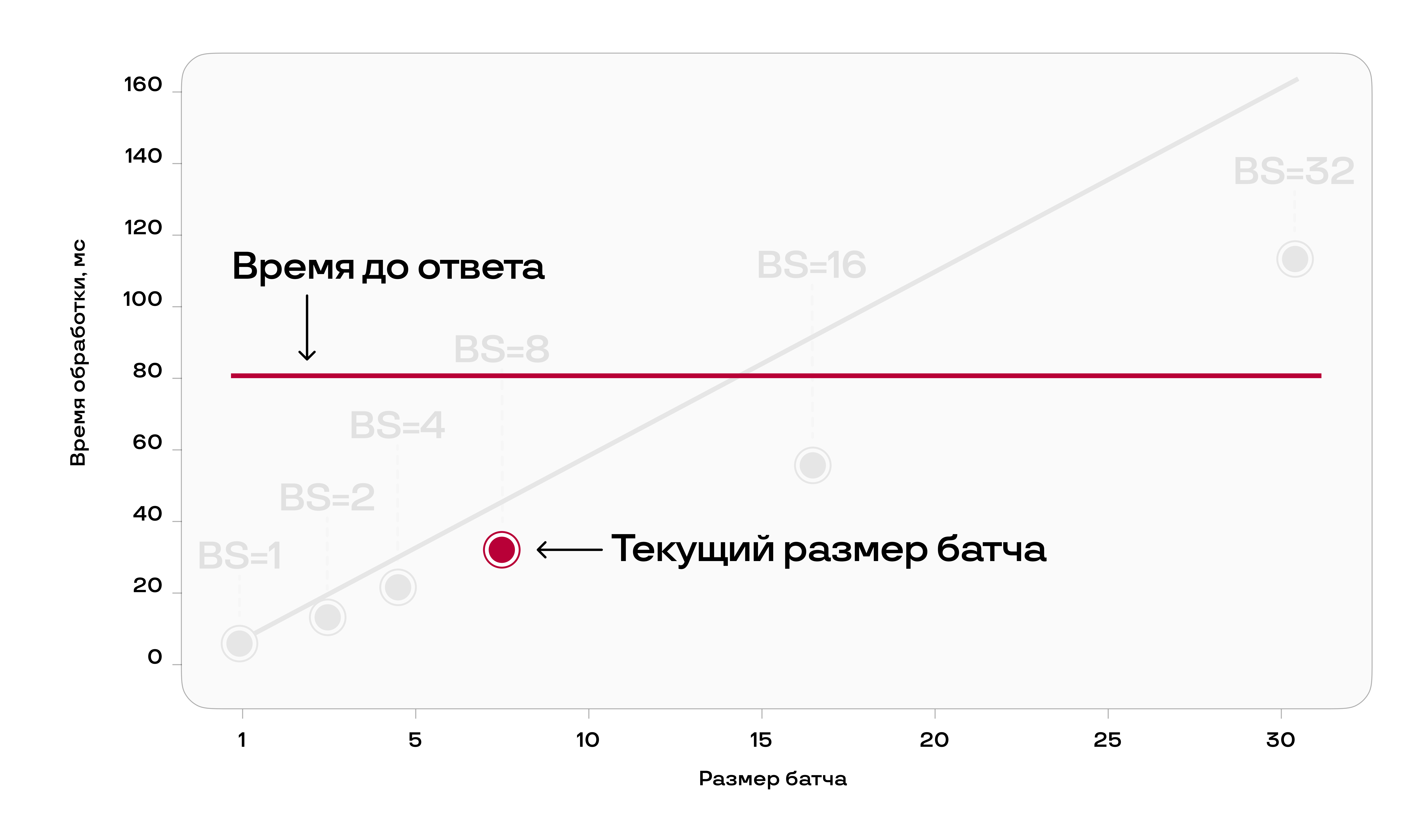


1. Оптимальные параметры
2. Динамический батч
3. Квантизация

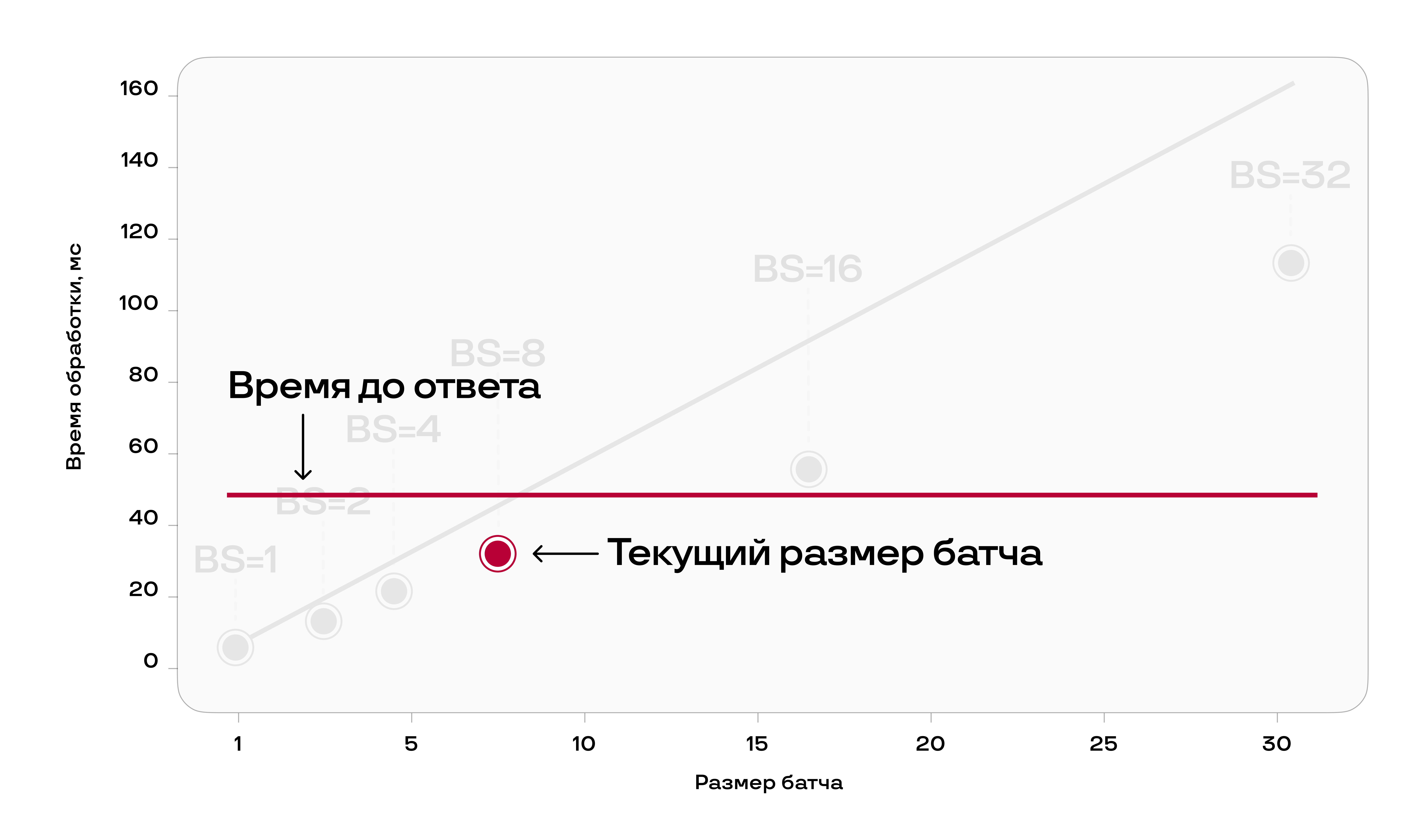
Оптимизация инференса: динамический батч



Оптимизация инференса: динамический батч



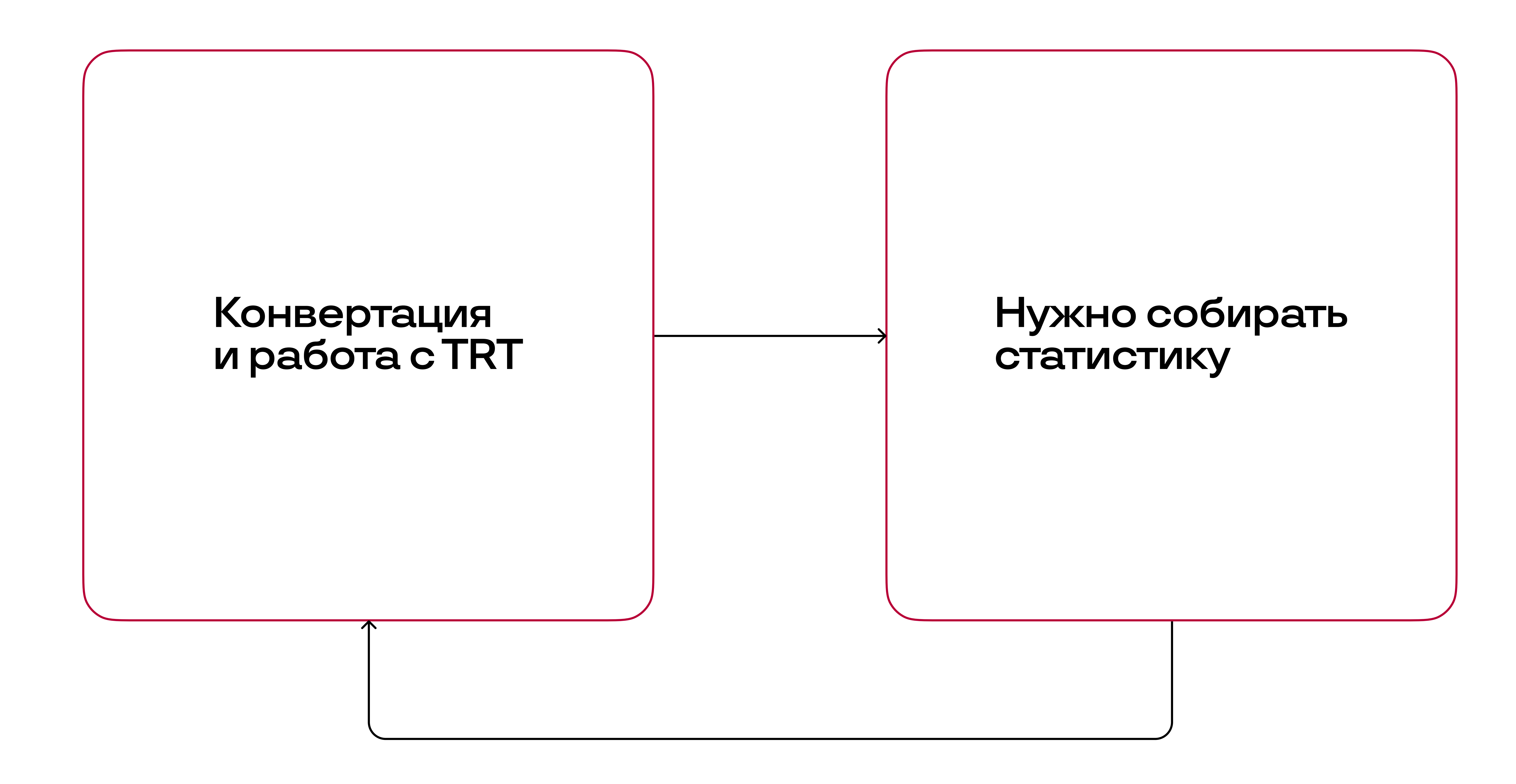
Оптимизация инференса: динамический батч



Батчинг



Особенности динамического батчинга



Варианты оптимизации

Модель

Инференс и батчинг

Вспомогательный код

Варианты оптимизации

Модель

Инференс и батчинг

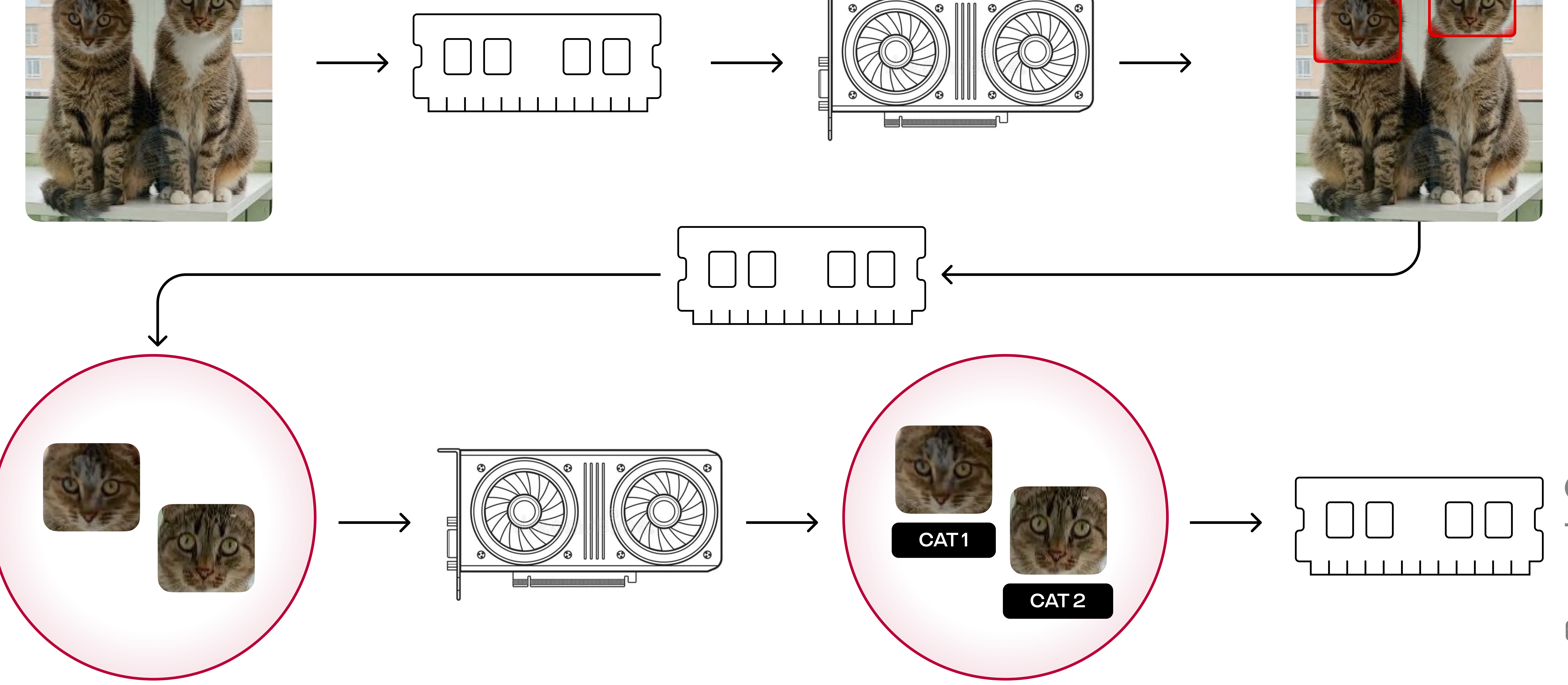
Вспомогательный код

Оптимизация кода: C++-модуль для Python

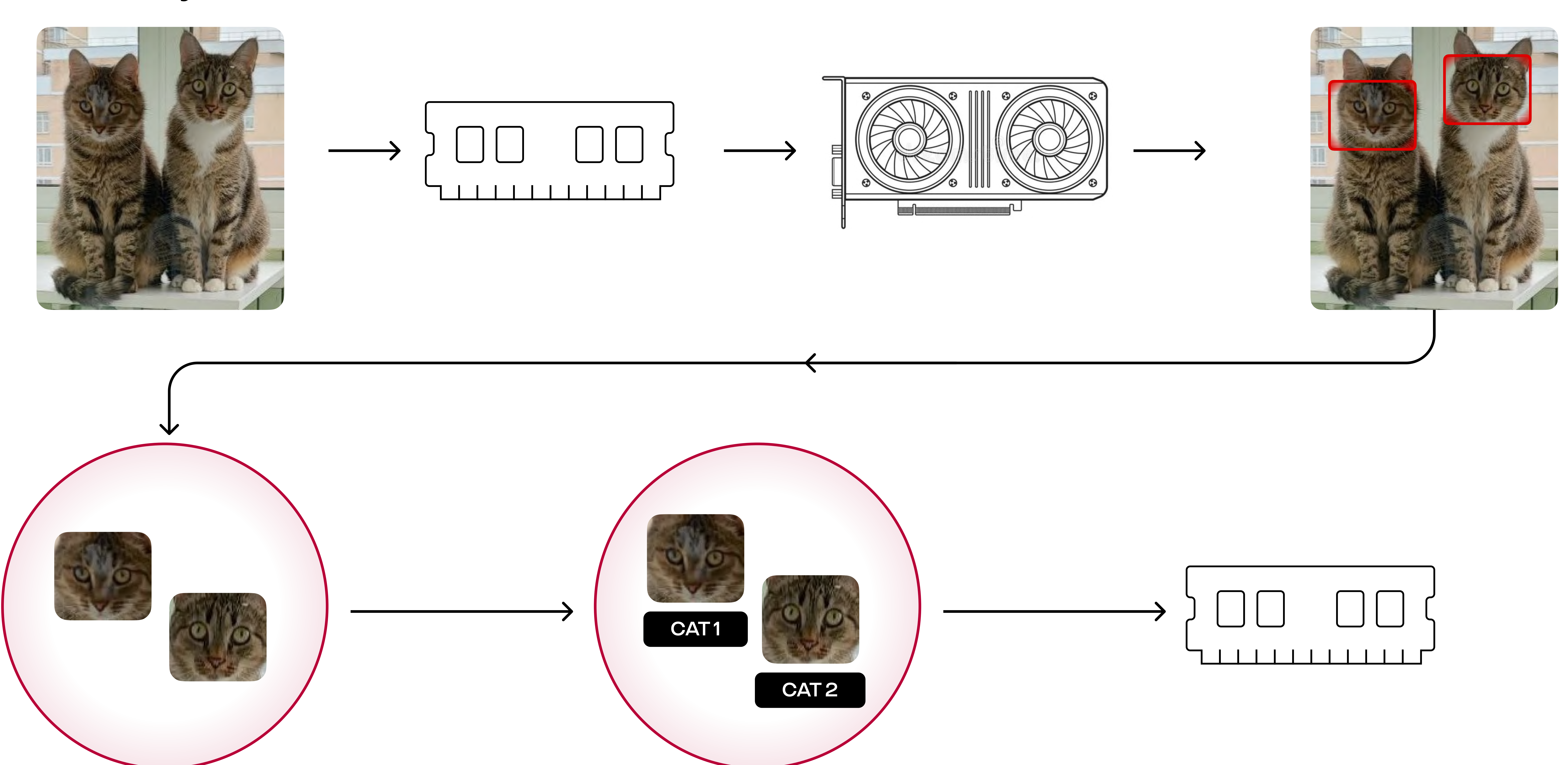
Все эти операции целиком выполняются на GPU:

1. JPEG-кодирование / декодирование
2. Любая обработка изображений на OpenCV: обрезка, экстракция масок, нормализация, и т.д.
3. Конвертация: OpenCV Mat <-> TRT / Torch Tensor
4. Инференс моделей на TensorRT
5. Центральный менеджер памяти GPU

Почему это важно?



Почему это важно?

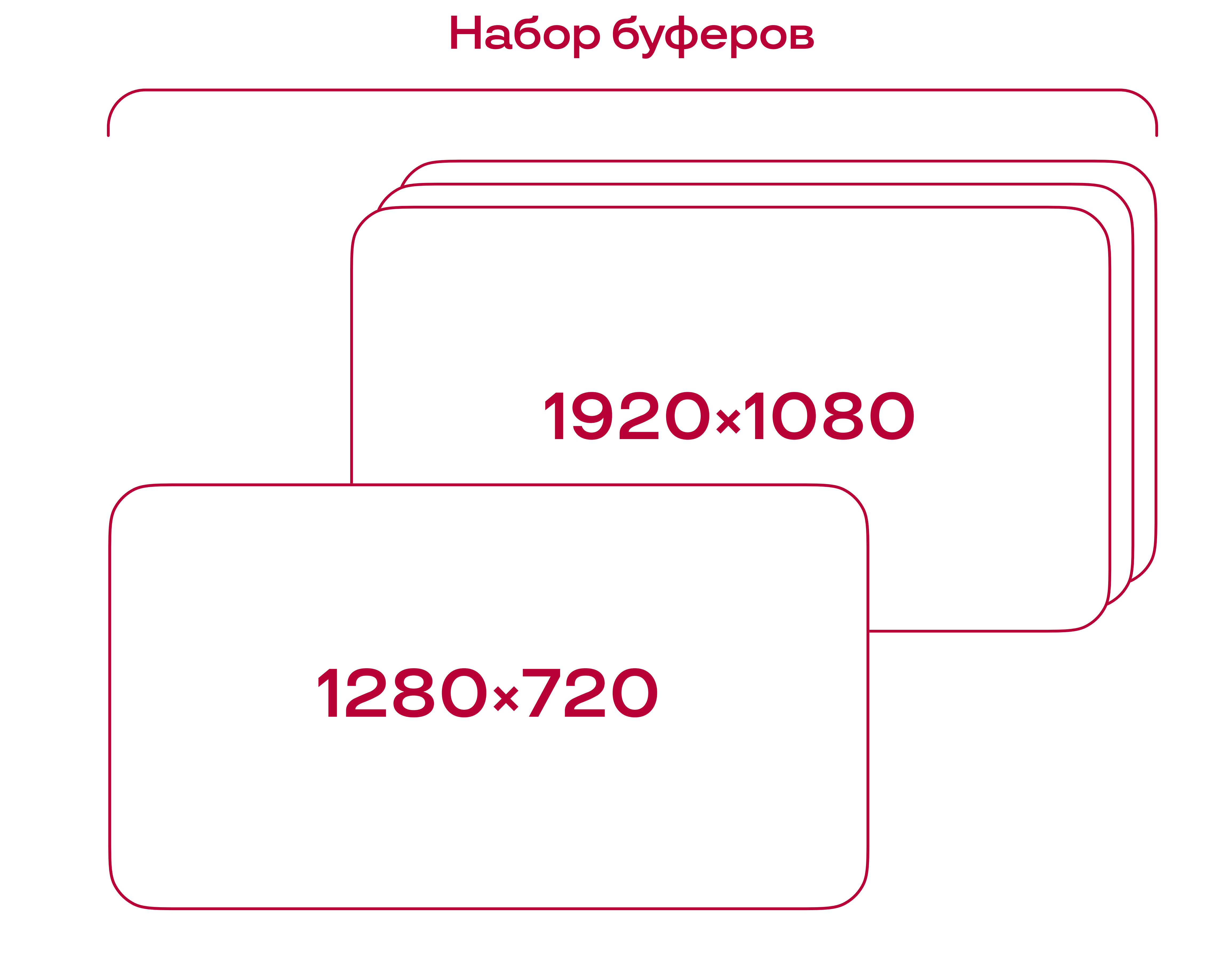


Лишние аллокации

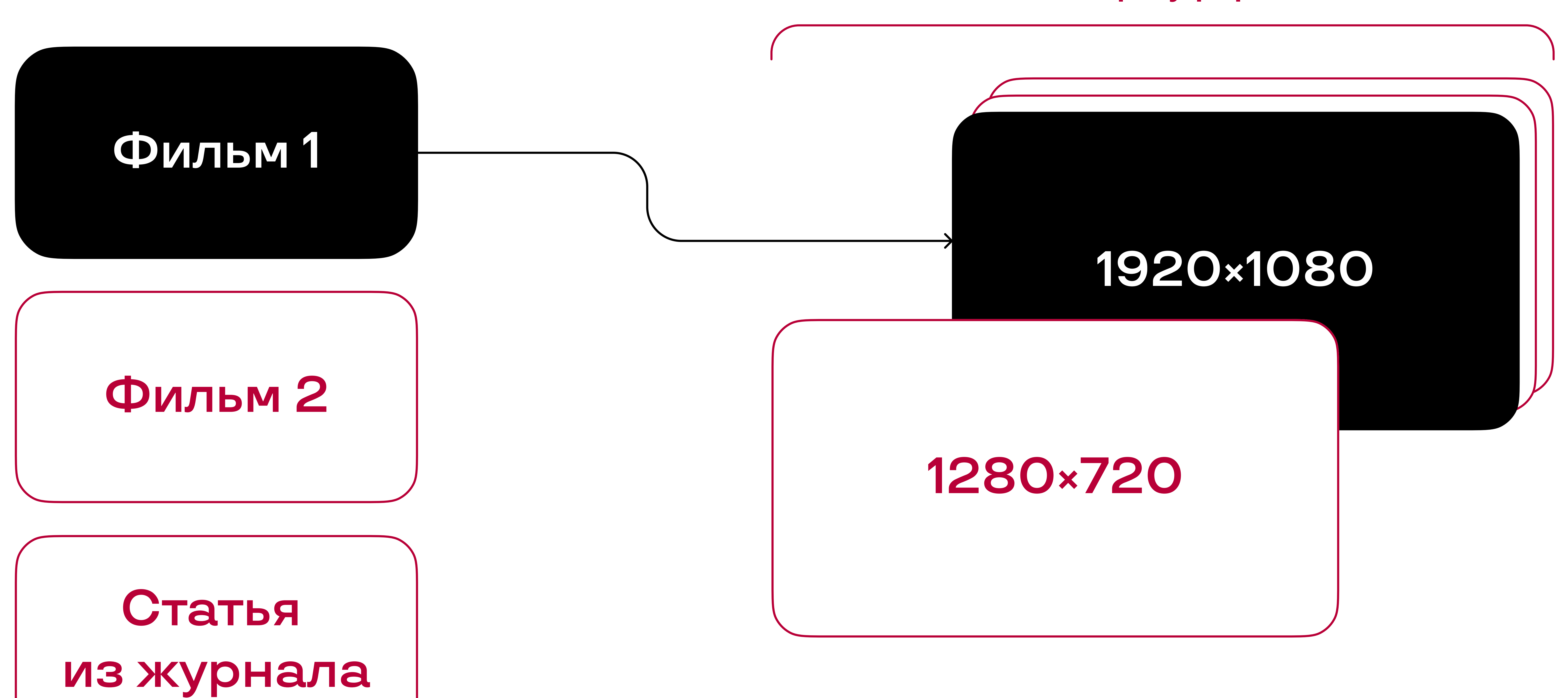
Фильм 1

Фильм 2

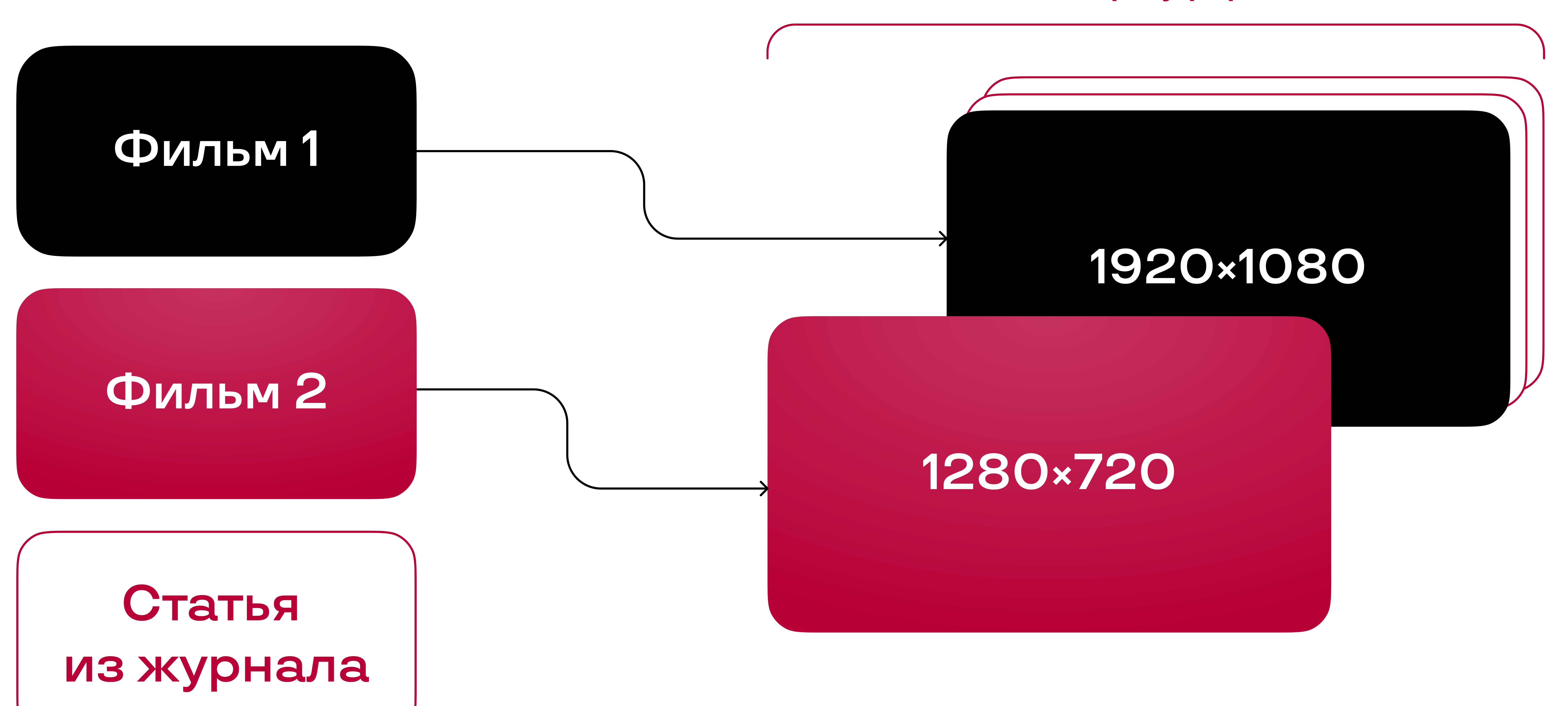
Статья
из журнала



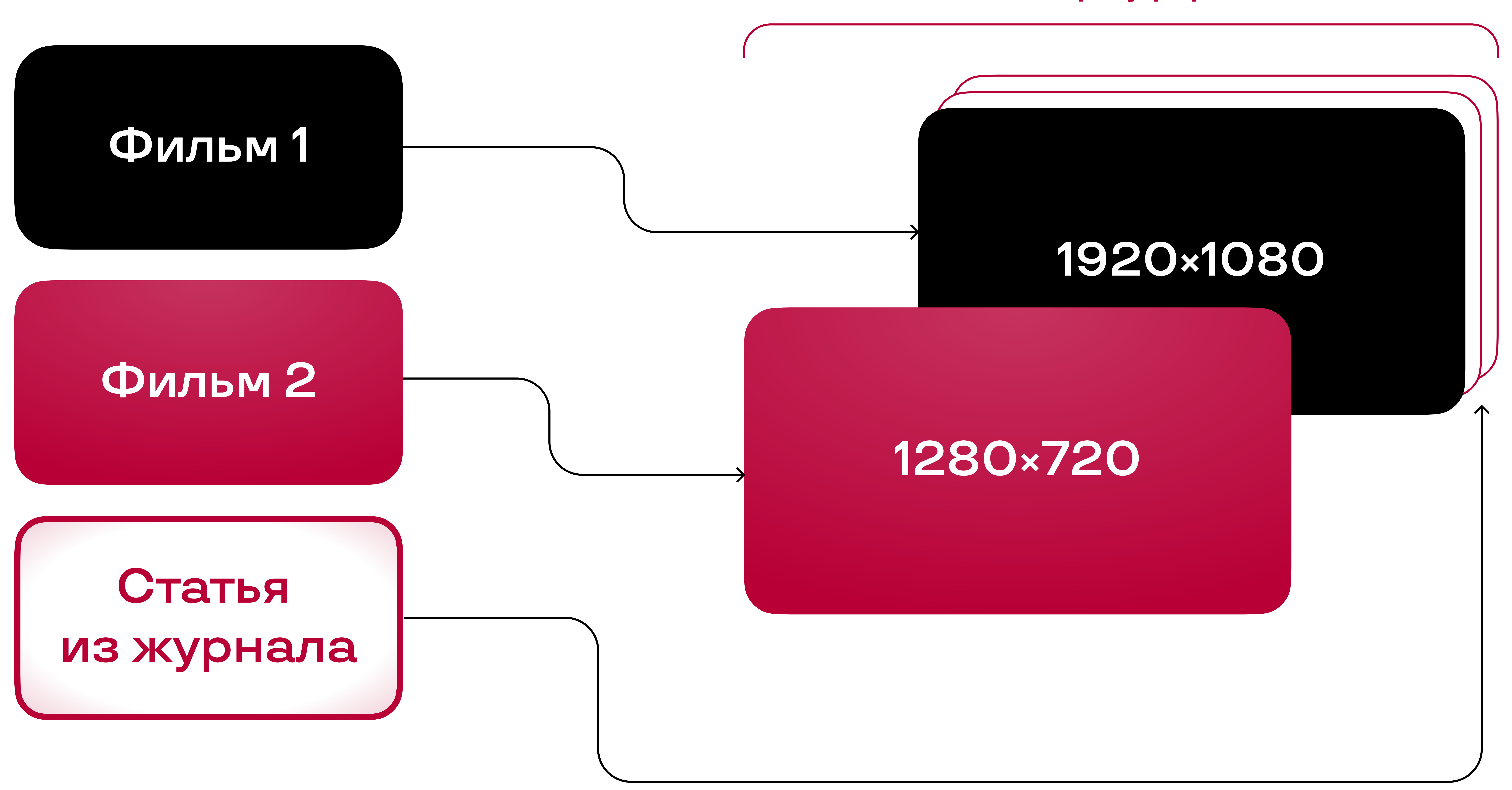
Лишние аллокации



Лишние аллокации



Лишние аллокации



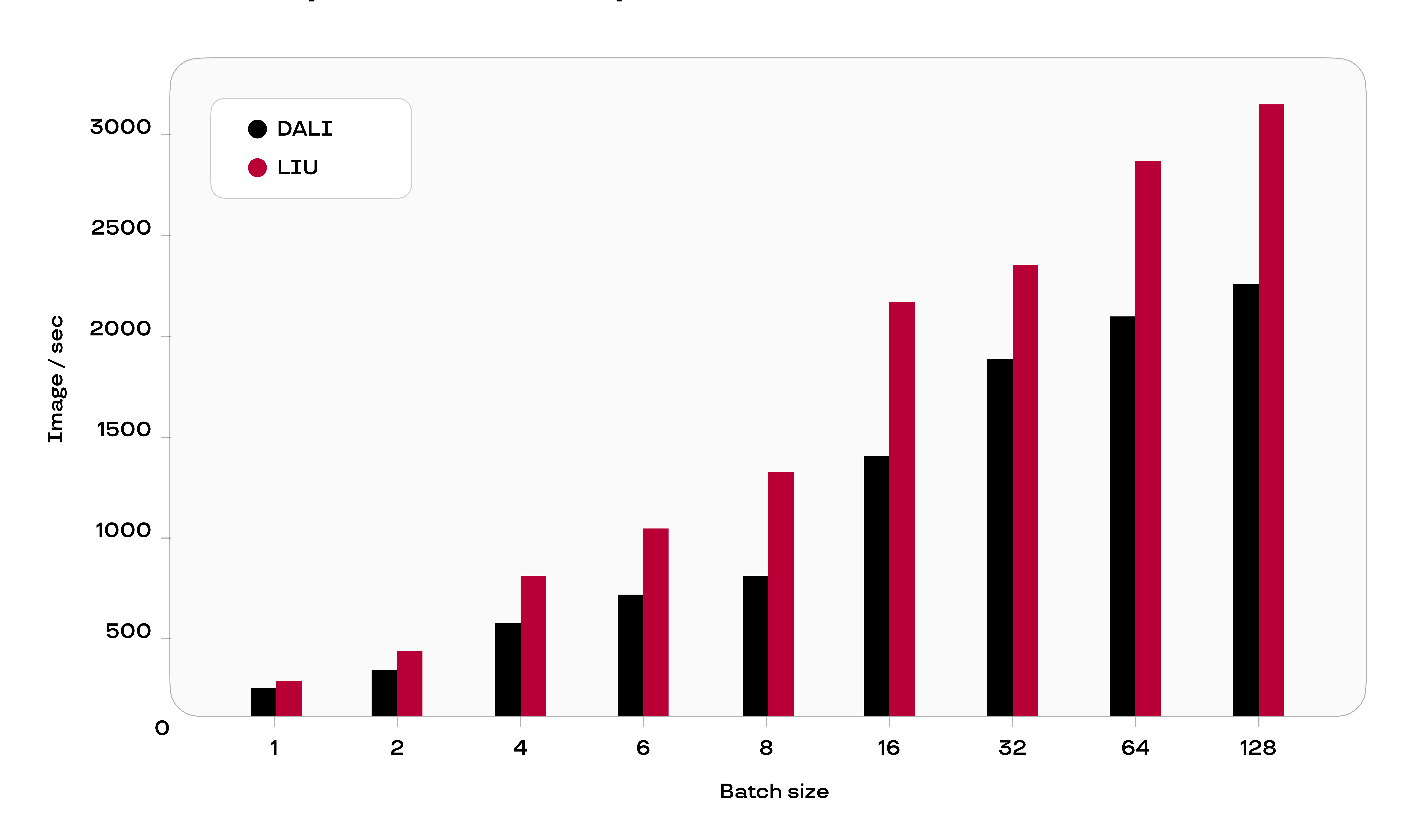
Почему не Nvidia Triton Inference Server

Triton Inference Server

VS

- Своя обработка ошибок
- Гибкость и свои интерфейсы
- Легковесность
- Скорость

Сравнение скоростей декодирования



Использование CPU

Перенос инференса и всех операций на GPU

Использование CPU

Перенос инференса и всех операций на GPU

→

Большая загрузка карты при «простаивании» CPU

Использование CPU

Перенос инференса и всех операций на GPU

→

Большая загрузка карты при «простаивании» CPU

Передача части операций CPU

Использование CPU

Перенос инференса и всех операций на GPU	→	Большая загрузка карты при «простаивании» CPU
Передача части операций CPU	→	Более равномерная загрузка девайсов

Использование CPU

Перенос инференса и всех операций на GPU → Большая загрузка карты при «простаивании» CPU

Передача части операций CPU → Более равномерная загрузка девайсов

Чем можно загрузить CPU?

1. Распаковка png и webp
2. Попытка восстановления битых jpeg
3. Гибридный режим для декодирования

1 Совсем другой язык: Julia, Golang, Rust, C++

О языках

1. Совсем другой язык: Julia, Golang, Rust, C++
2. Почти питон: Numba, Cython

О языках

1. Совсем другой язык: Julia, Golang, Rust, C++
2. Почти питон: Numba, Cython
3. Модули на C++, основной Python

Варианты оптимизации

Модель

Инференс и батчинг

Вспомогательный код

Варианты оптимизации



Варианты оптимизации



Бонус: ускорение тренировок моделей в ~ 2 раза

Спасибо за внимание!

Оценить доклад

Telegram
@Grigoriy_Alekseenko

Канал команды в Telegram
t.me/layercv



